Predictive Diabetes Insights: A Statistical Approach

Statistical Learning
2022/2023

29/05/2023

Università degli Studi di Padova

Dipartimento Matematica

Francesco Pittorino	2090920
Giulio Nebbiai	2092296

Purpose of the study

- Aim:

Binary classification for the recognition of the presence of diabetes in the population consisting of females at least 21 years old having Pima Indian heritage:

1. Estimation of a model capable of forecasting the condition
2. Detect the most important features for the purpose

- Data:

Diabetes dataset, which is made up by 768 examples described by 9 variables.

Data features

Predictors:

- Pregnancies
- Glucose
- BloodPressure
- SkinThickness
- Insulin
- BMI
- DiabetesPedigreeFunction
- Age

Response Variable:

- Outcome

Data preprocessing and filtering

- Check the presence of NA values:

1. SkinThickness
2. BMI
3. BloodPressure

Predictors:

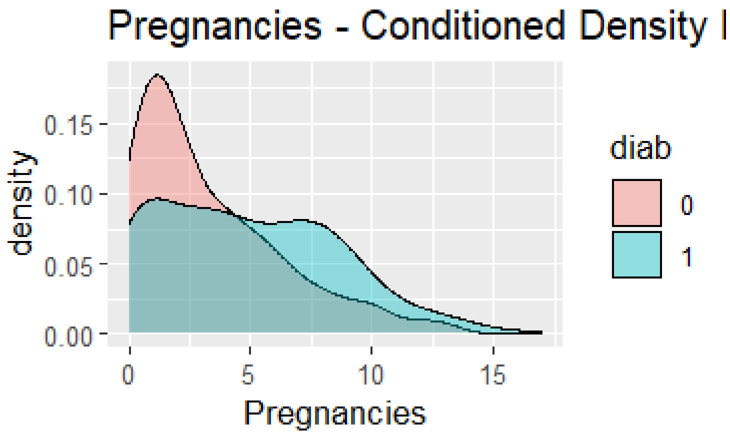
- Pregnancies
- Glucose
- BloodPressure
- ~~SkinThickness~~
- Insulin
- BMI
- DiabetesPedigreeFunction
- Age

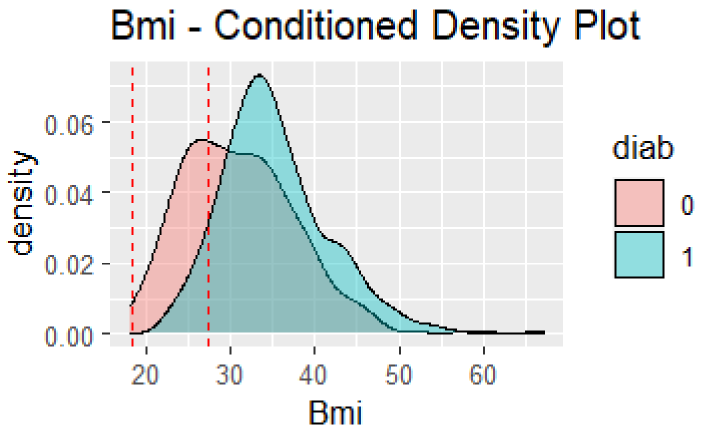
Response Variable:

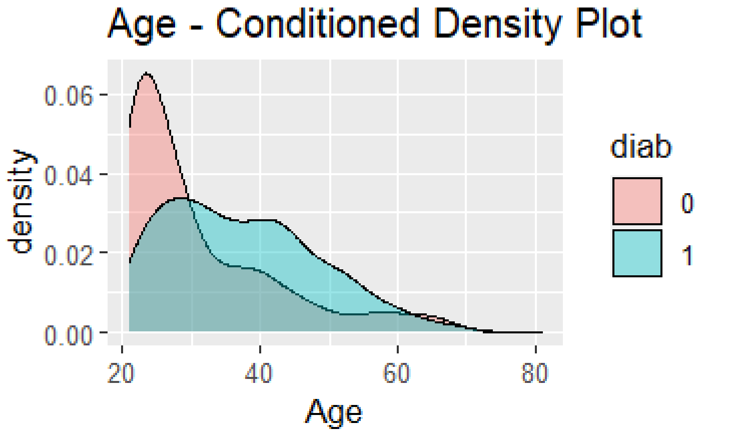
- Outcome

Data: 729 observations described by 8 variables

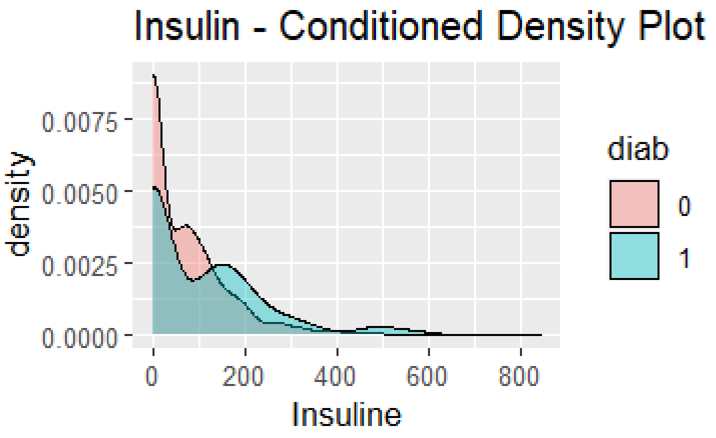
Conditioned distributions: demographic features


Pregnancies - Conditioned Density I

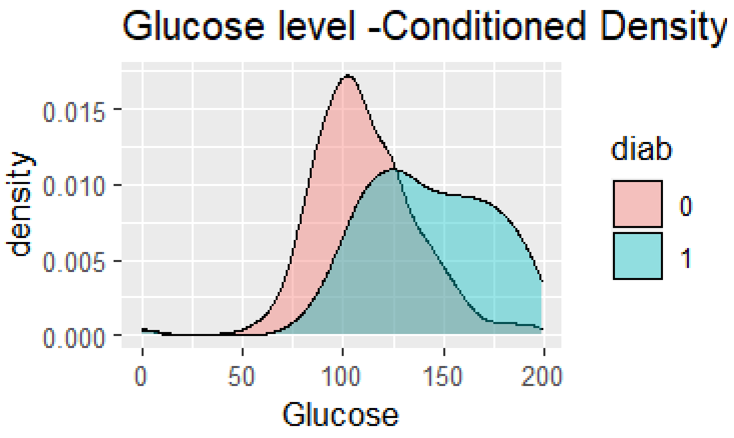

Bmi - Conditioned Density Plot


Age - Conditioned Density Plot

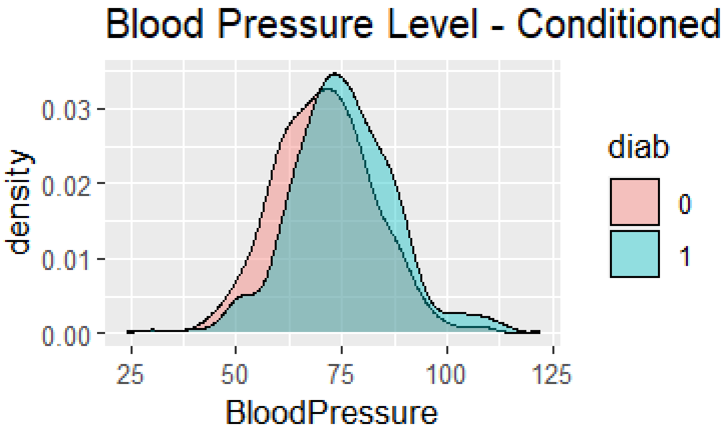
Conditioned distributions: medical condition features



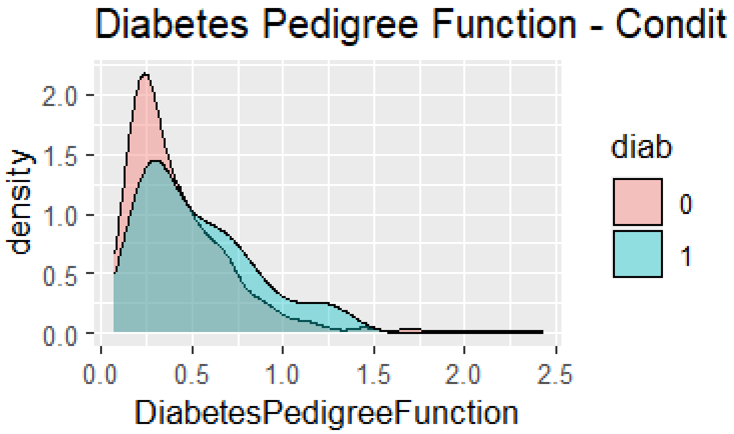
Insulin - Conditioned Density Plot



Glucose level -Conditioned Density

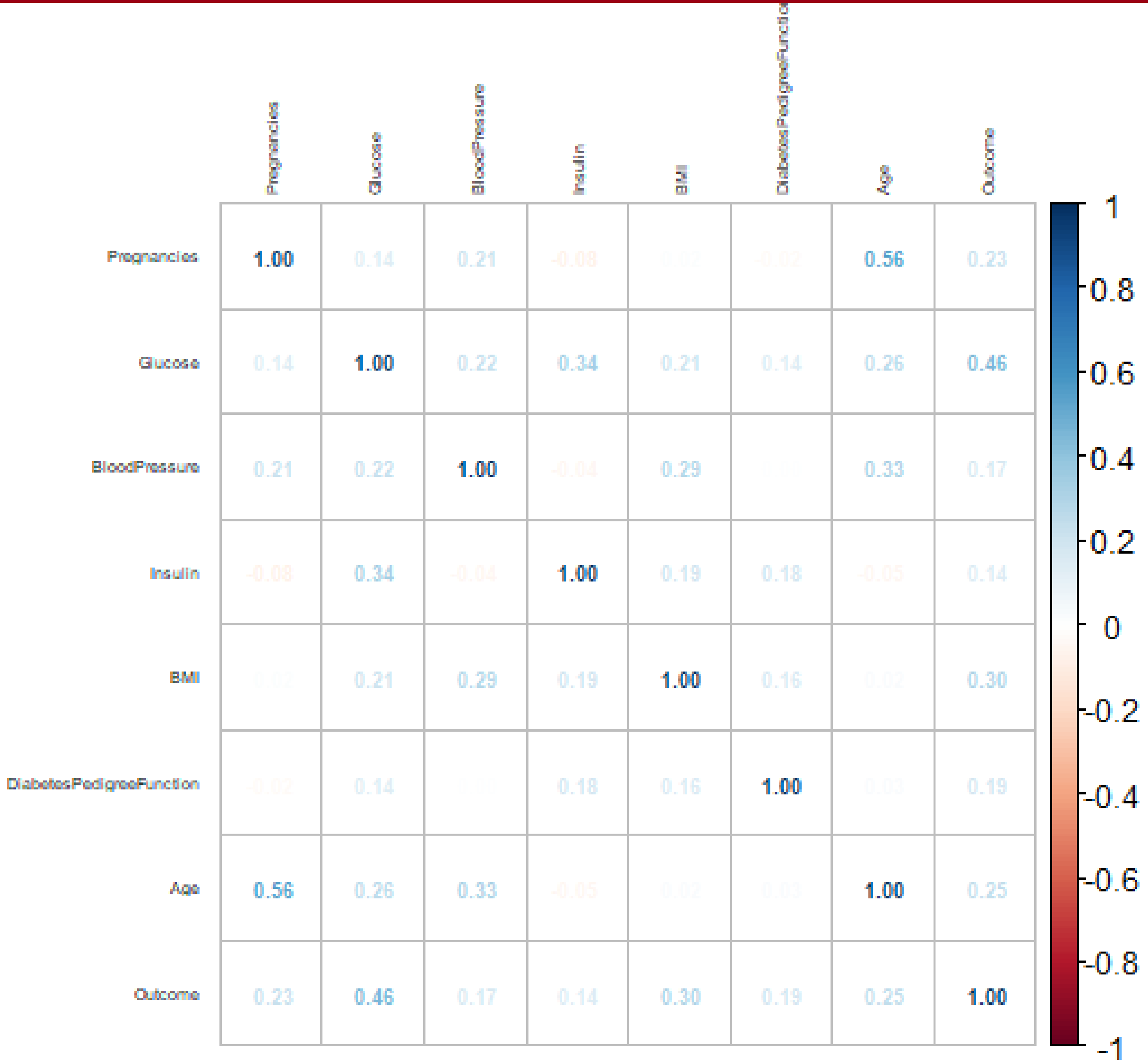


Blood Pressure Level - Conditioned



Diabetes Pedigree Function - Condit

Correlations



	Pregnancies	Glucose	BloodPressure	Insulin	BMI	DiabetesPedigreeFunctio	Age	Outcome
Pregnancies	1.00	0.14	0.21	-0.08	0.02	-0.02	0.56	0.23
Glucose	0.14	1.00	0.22	0.34	0.21	0.14	0.26	0.46
BloodPressure	0.21	0.22	1.00	-0.04	0.29	0.00	0.33	0.17
Insulin	-0.08	0.34	-0.04	1.00	0.19	0.18	-0.05	0.14
BMI	0.02	0.21	0.29	0.19	1.00	0.16	0.02	0.30
DiabetesPedigreeFunction	-0.02	0.14	0.00	0.18	0.16	1.00	0.03	0.19
Age	0.56	0.26	0.33	-0.05	0.02	0.03	1.00	0.25
Outcome	0.23	0.46	0.17	0.14	0.30	0.19	0.25	1.00

Dataset splitting

In order to evaluate the performances of our models, we split randomly the dataset into:

- Training set: 547 examples (75% of the dataset)
- Test set: 152 examples (25% of the dataset)

Metrics

To permit a comparison between the models, the following metrics are taken into account:

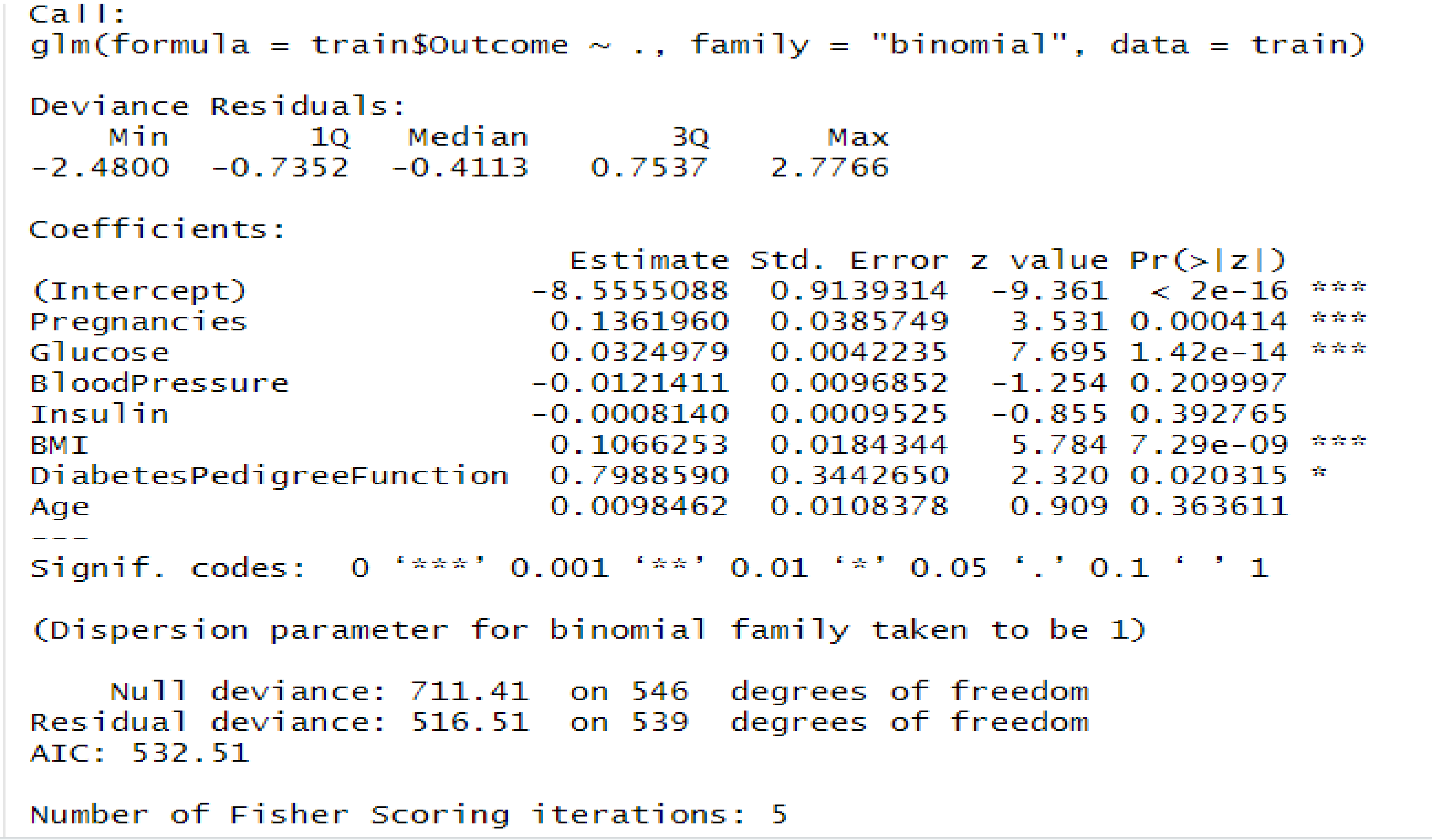
- Accuracy:

$$\frac{True\ Positives + True\ Negatives}{True\ Positives + False\ Positives + False\ Negatives + True\ Negatives}$$

- Sensitivity:

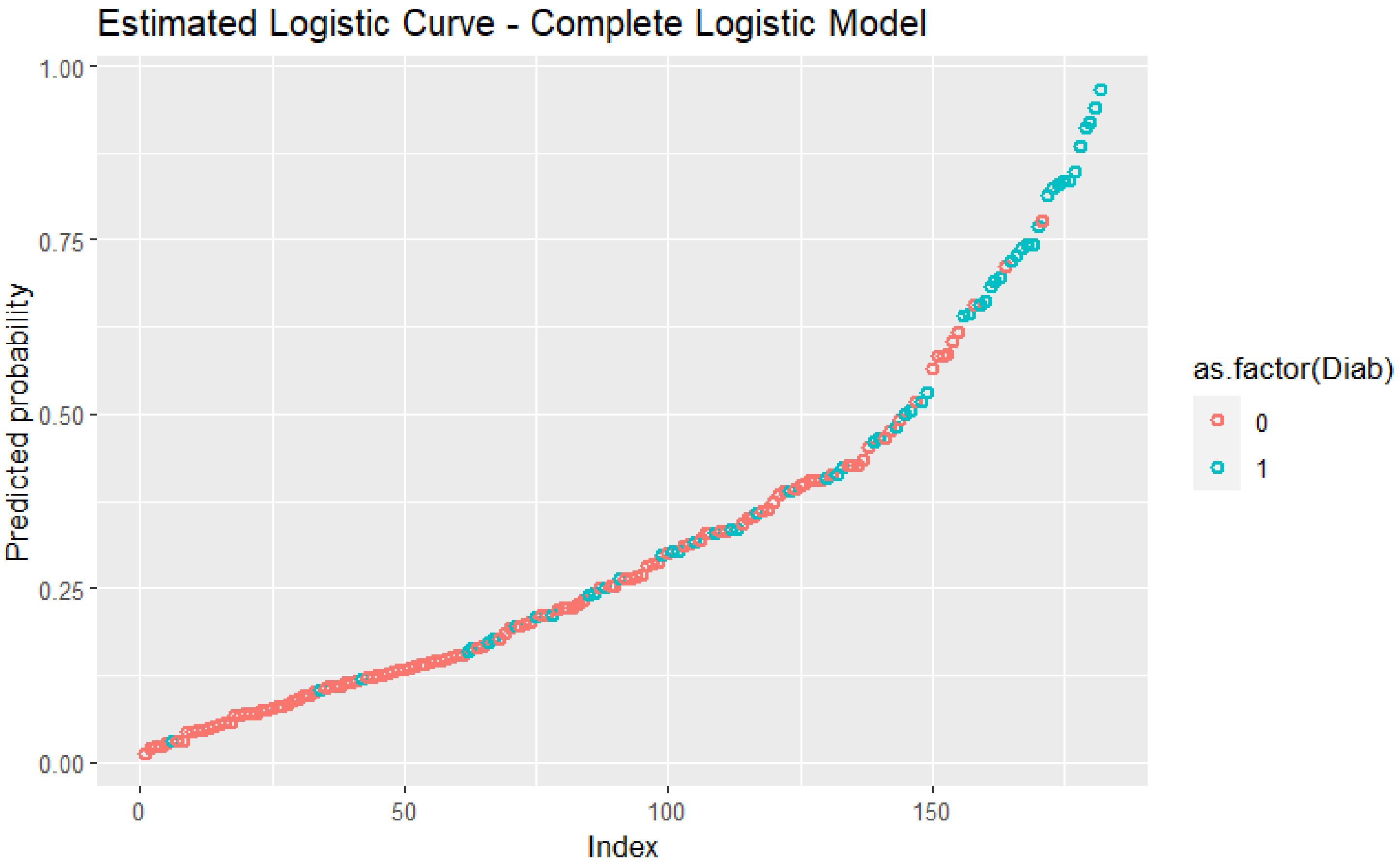
$$\frac{True\ Positives}{True\ Positives\ +\ False\ Negatives}$$

Complete Logistic model



```
Call:
glm(formula = train$Outcome ~ ., family = "binomial", data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.4800  -0.7352  -0.4113   0.7537   2.7766  

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -8.5555088  0.9139314  -9.361  < 2e-16 ***
Pregnancies               0.1361960  0.0385749   3.531 0.000414 ***
Glucose                   0.0324979  0.0042235   7.695 1.42e-14 ***
BloodPressure            -0.0121411  0.0096852  -1.254 0.209997    
Insulin                  -0.0008140  0.0009525  -0.855 0.392765    
BMI                       0.1066253  0.0184344   5.784 7.29e-09 ***
DiabetesPedigreeFunction  0.7988590  0.3442650   2.320 0.020315 *  
Age                       0.0098462  0.0108378   0.909 0.363611    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 711.41  on 546  degrees of freedom
Residual deviance: 516.51  on 539  degrees of freedom
AIC: 532.51

Number of Fisher Scoring iterations: 5
```

Complete Logistic model



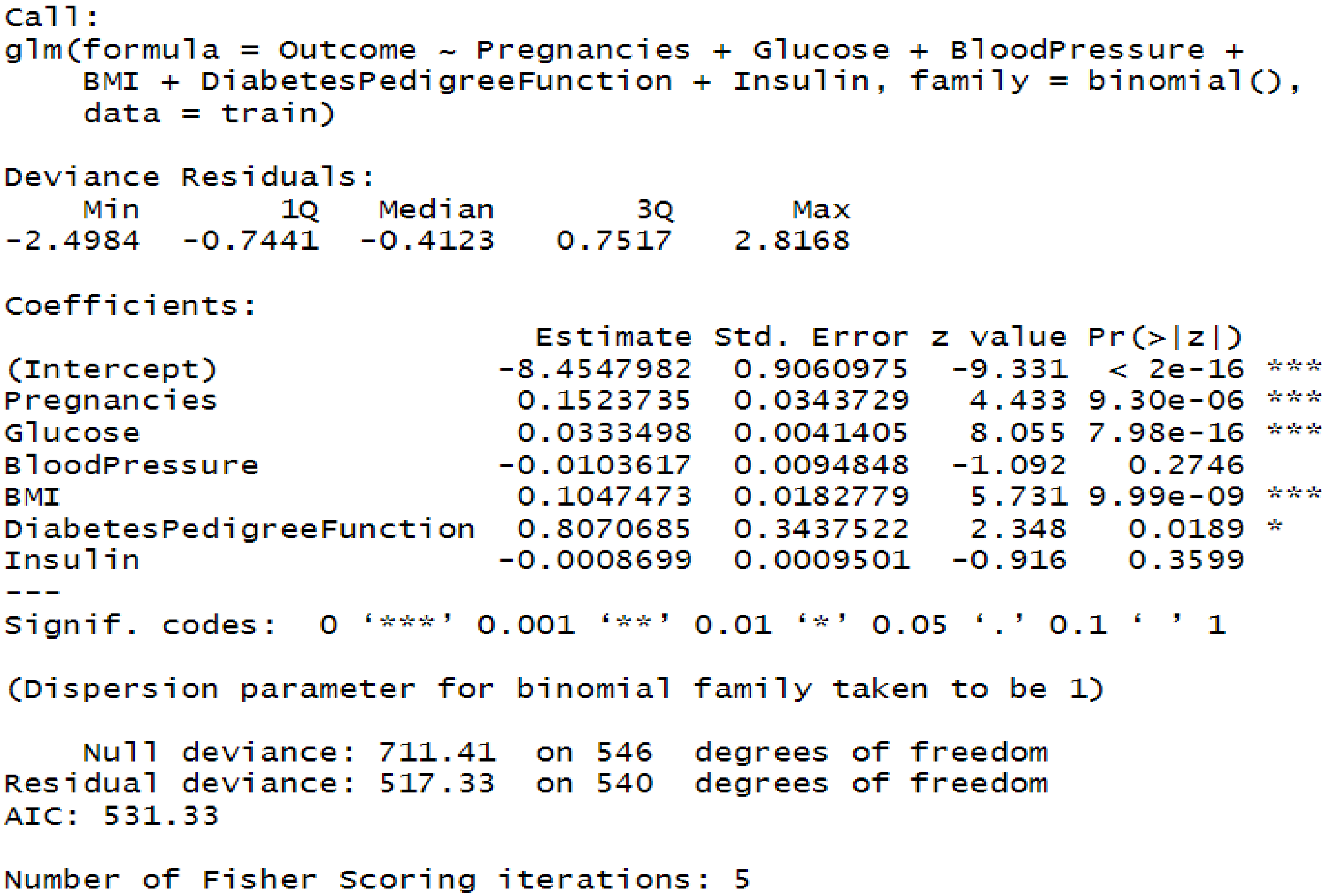
METRIC	VALUE
Accuracy	0.6923077
Sensitivity	0.7368421

Multicollinearity Check: VIF

Predictors	VIF value
Pregnancies	1.397697
Glucose	1.256495
BloodPressure	1.265047
Insulin	1.246988
BMI	1.155411
DiabetesPedigreeFunction	1.020217
Age	1.518657

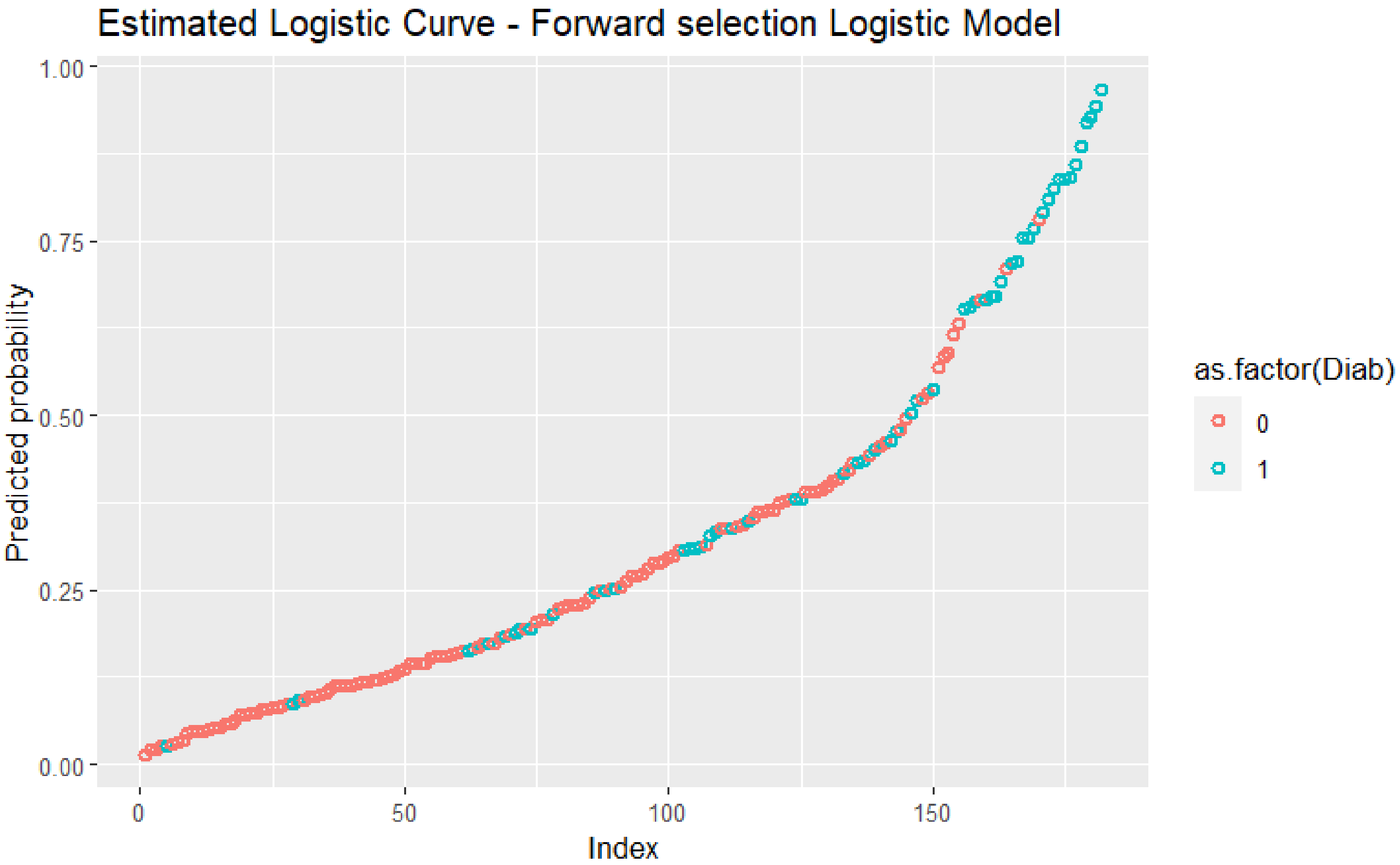
Threshold (1/(1-R^2)	1.377341

Forward Selection Logistic Regression



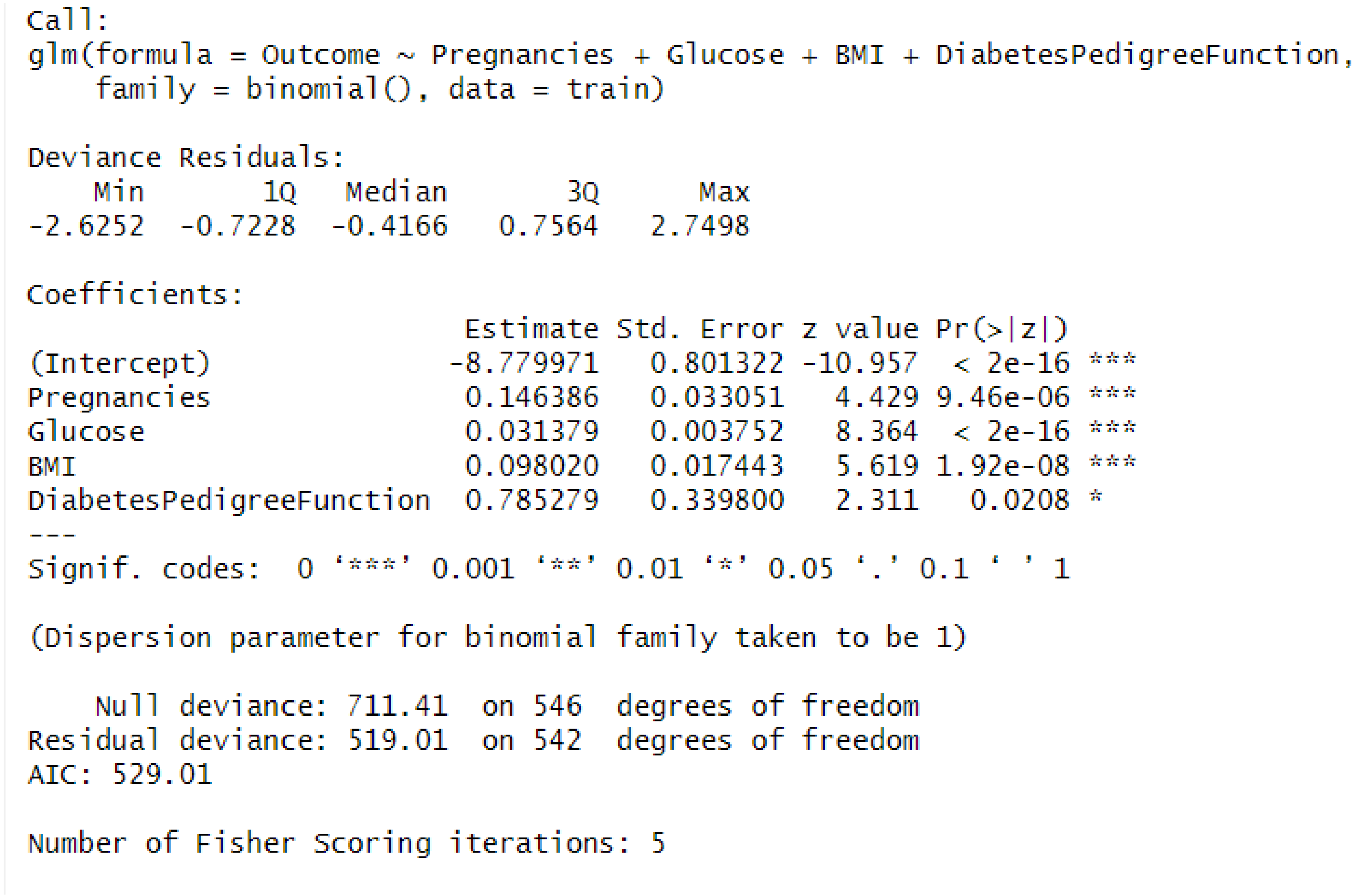
```
Call:
glm(formula = Outcome ~ Pregnancies + Glucose + BloodPressure +
    BMI + DiabetesPedigreeFunction + Insulin, family = binomial(),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4984  -0.7441  -0.4123   0.7517   2.8168

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -8.4547982  0.9060975  -9.331  < 2e-16 ***
Pregnancies               0.1523735  0.0343729   4.433 9.30e-06 ***
Glucose                   0.0333498  0.0041405   8.055 7.98e-16 ***
BloodPressure            -0.0103617  0.0094848  -1.092   0.2746
BMI                       0.1047473  0.0182779   5.731 9.99e-09 ***
DiabetesPedigreeFunction  0.8070685  0.3437522   2.348   0.0189 *
Insulin                  -0.0008699  0.0009501  -0.916   0.3599
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 711.41  on 546  degrees of freedom
Residual deviance: 517.33  on 540  degrees of freedom
AIC: 531.33

Number of Fisher Scoring iterations: 5
```

Forward Selection Logistic Regression



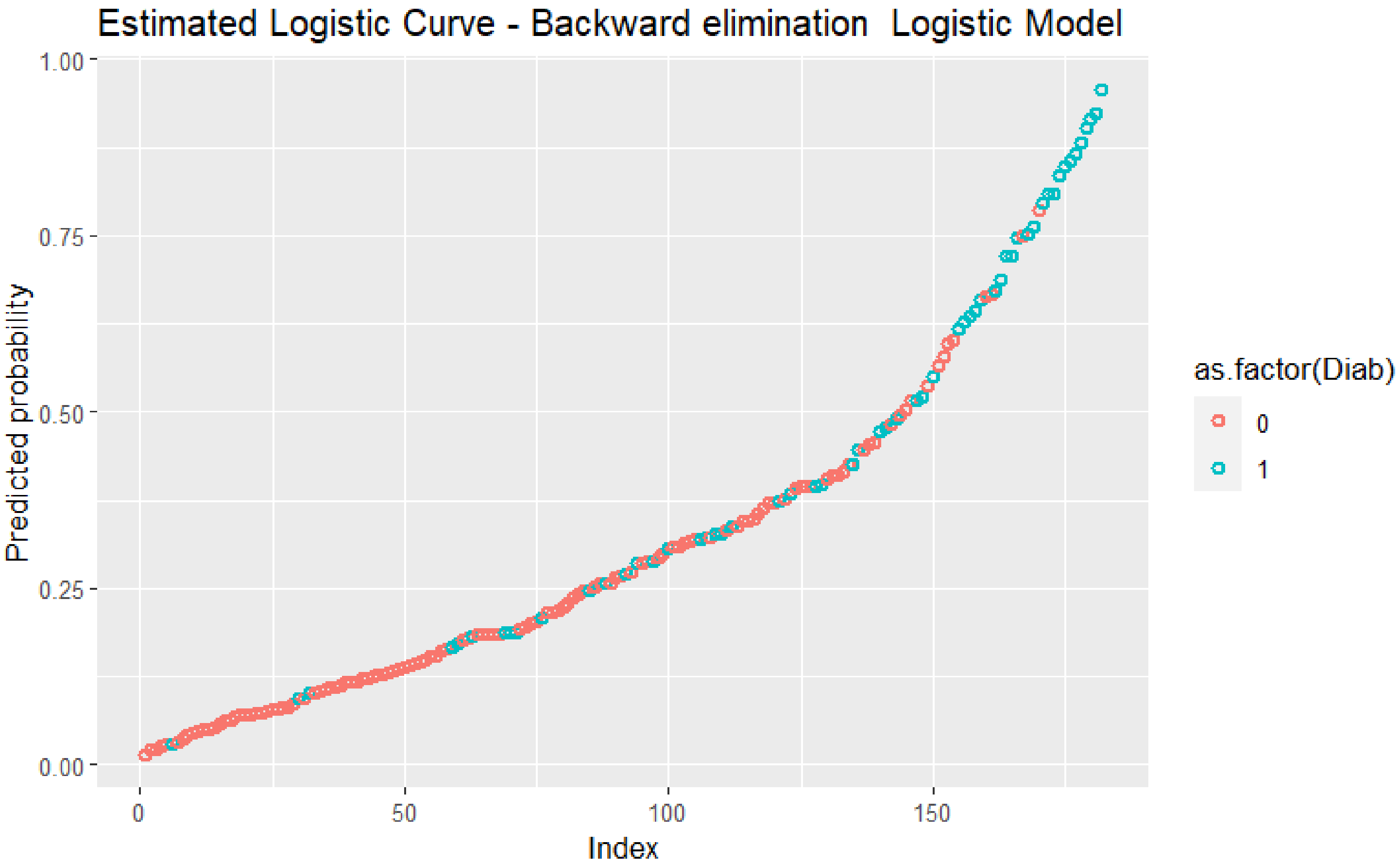
METRIC	VALUE
Accuracy	0.7142857
Sensitivity	0.754386

Backward Elimination Logistic Regression



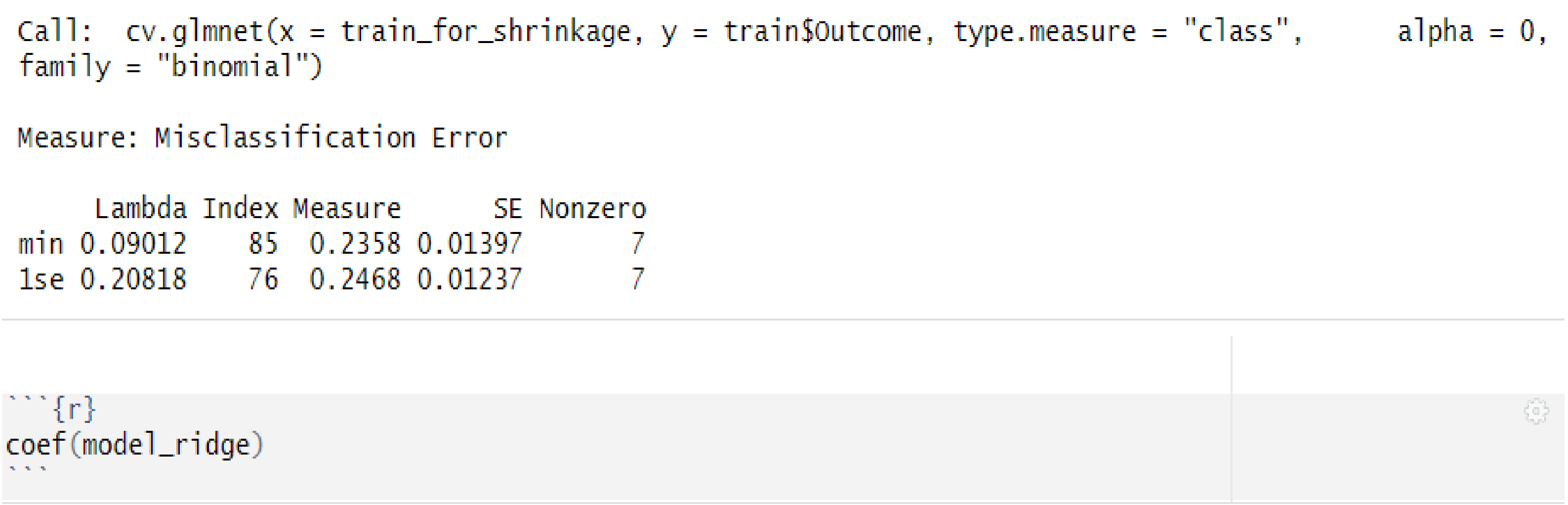
```
Call:
glm(formula = Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction,
    family = binomial(), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6252  -0.7228  -0.4166   0.7564   2.7498

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -8.779971   0.801322 -10.957  < 2e-16 ***
Pregnancies               0.146386   0.033051   4.429 9.46e-06 ***
Glucose                   0.031379   0.003752   8.364  < 2e-16 ***
BMI                       0.098020   0.017443   5.619 1.92e-08 ***
DiabetesPedigreeFunction  0.785279   0.339800   2.311   0.0208 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 711.41  on 546  degrees of freedom
Residual deviance: 519.01  on 542  degrees of freedom
AIC: 529.01

Number of Fisher Scoring iterations: 5
```

Backward Elimination Logistic Regression



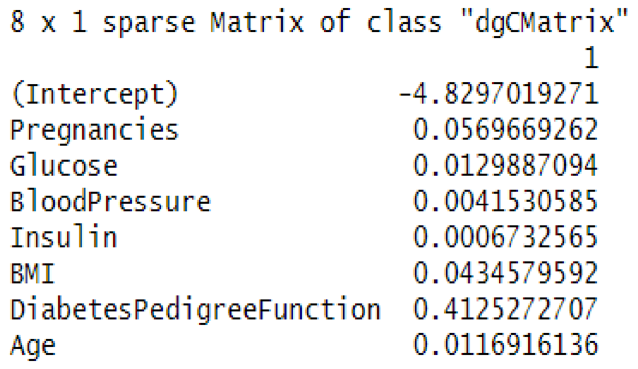
METRIC	VALUE
Accuracy	0.6923077
Sensitivity	0.7368421

Shrinkage methods – Ridge Regression



```
Call:  cv.glmnet(x = train_for_shrinkage, y = train$Outcome, type.measure = "class",      alpha = 0,
family = "binomial")

Measure: Misclassification Error

     Lambda Index Measure      SE Nonzero
min 0.09012    85   0.2358 0.01397       7
1se 0.20818    76   0.2468 0.01237       7
```

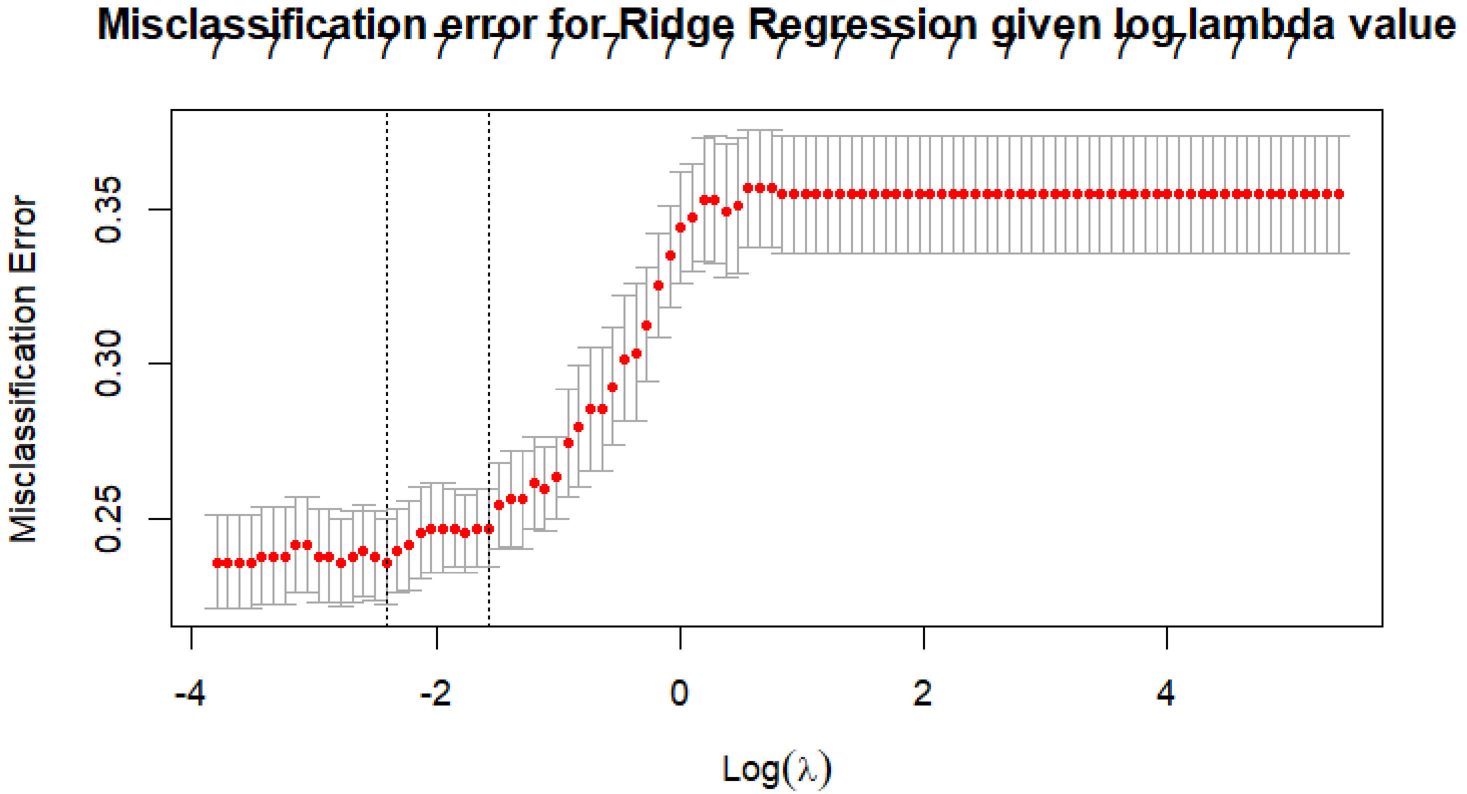
```{r}
coef(model_ridge)
```



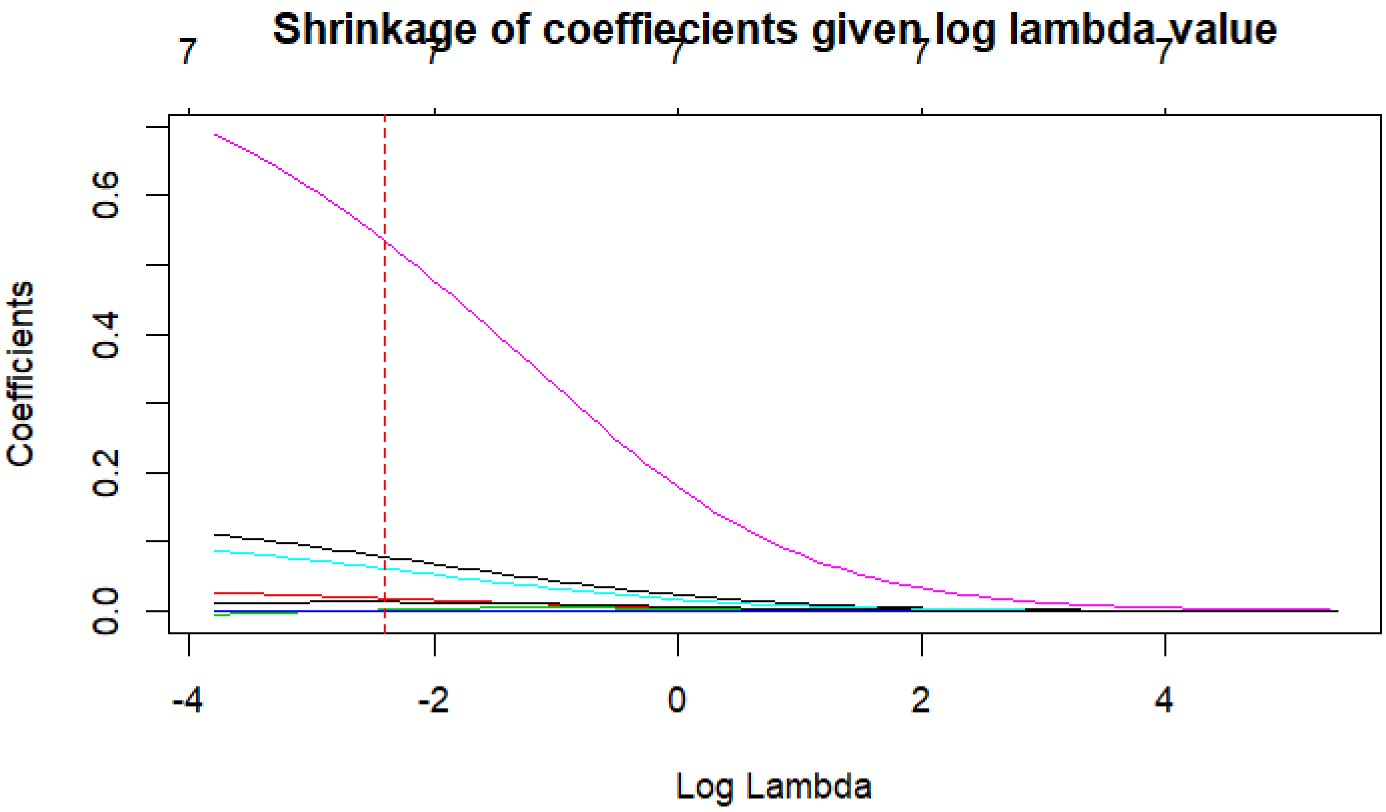
```
8 x 1 sparse Matrix of class "dgCMatrix"
                                       1
(Intercept)              -4.8297019271
Pregnancies               0.0569669262
Glucose                   0.0129887094
BloodPressure             0.0041530585
Insulin                   0.0006732565
BMI                       0.0434579592
DiabetesPedigreeFunction  0.4125272707
Age                       0.0116916136
```

METRIC	VALUE
Accuracy	0.6703297
Sensitivity	0.8245614

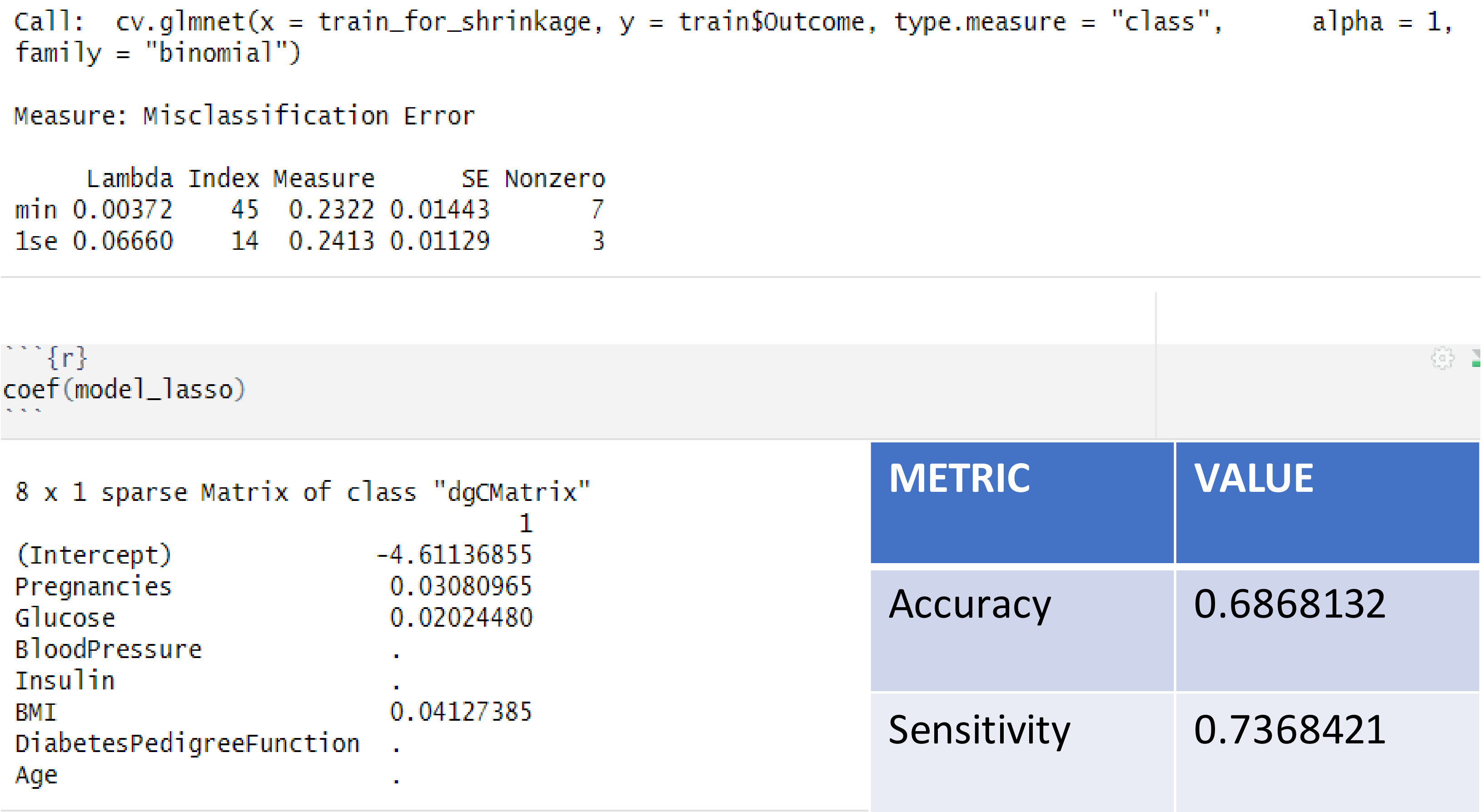
Ridge Regression



Ridge Regression



Shrinkage methods – Lasso Regression



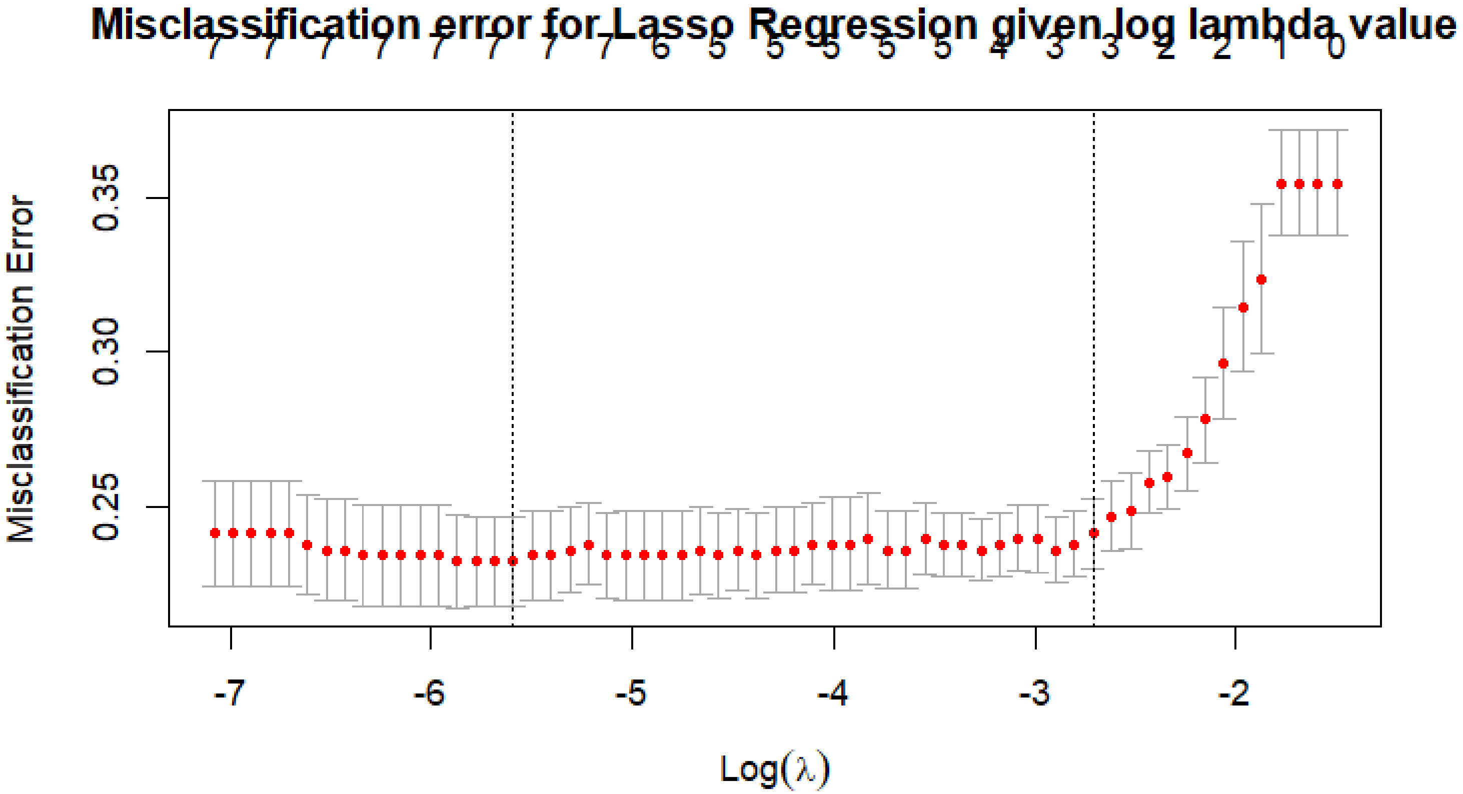
```
Call:  cv.glmnet(x = train_for_shrinkage, y = train$Outcome, type.measure = "class",      alpha = 1,
family = "binomial")

Measure: Misclassification Error

     Lambda Index Measure      SE Nonzero
min 0.00372    45   0.2322 0.01443       7
1se 0.06660    14   0.2413 0.01129       3
```

```{r}
coef(model_lasso)
```

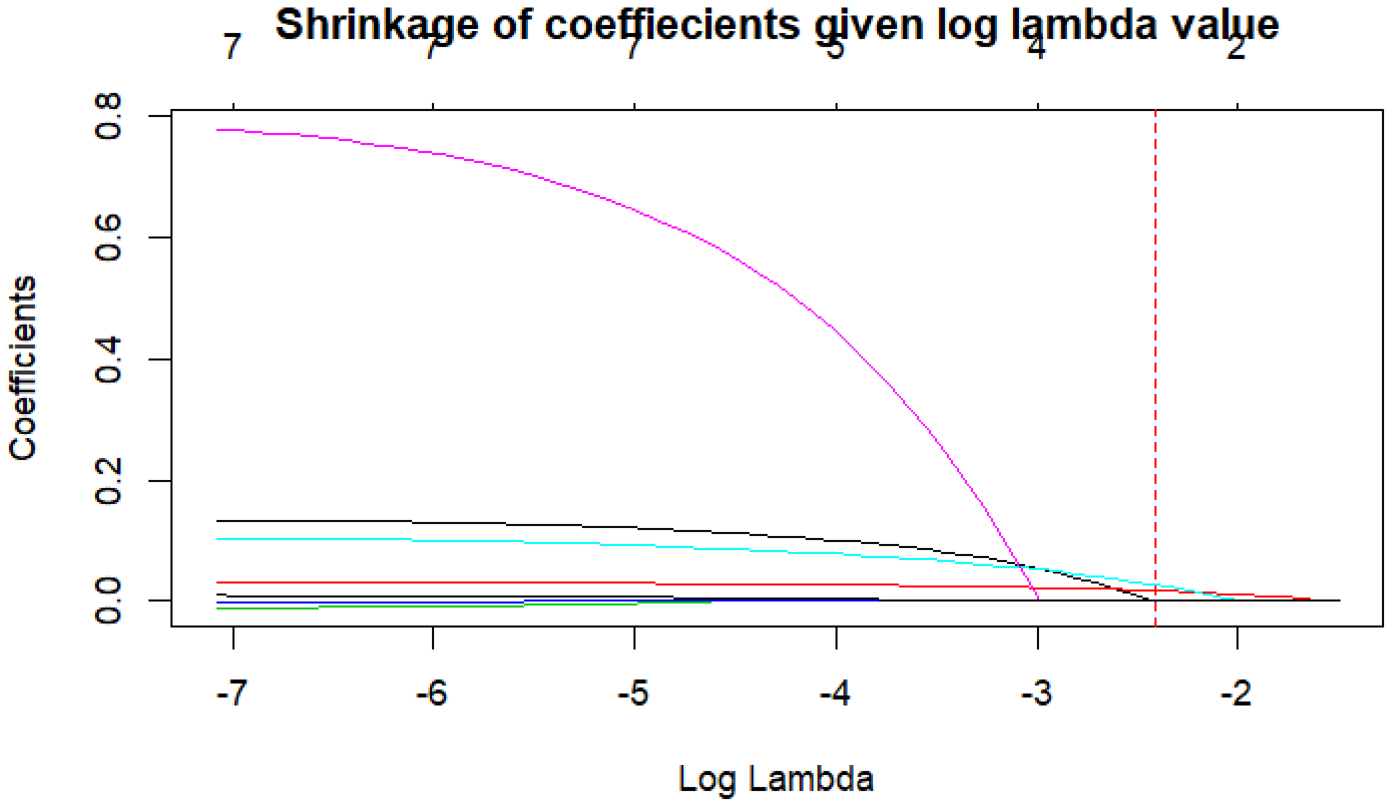
```
8 x 1 sparse Matrix of class "dgCMatrix"
                                  1
(Intercept)              -4.61136855
Pregnancies               0.03080965
Glucose                   0.02024480
BloodPressure             .
Insulin                   .
BMI                       0.04127385
DiabetesPedigreeFunction  .
Age                       .
```

METRIC	VALUE
Accuracy	0.6868132
Sensitivity	0.7368421

Lasso Regression



Lasso Regression

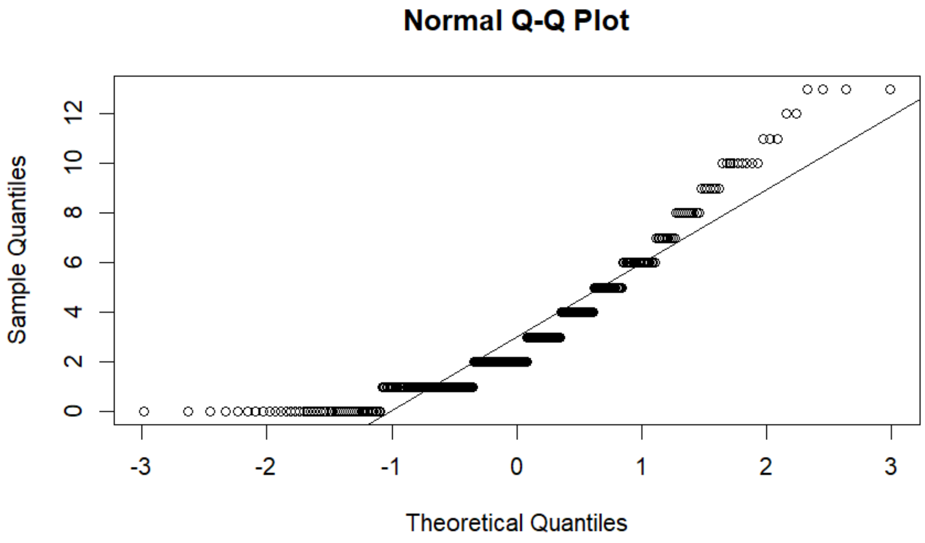


Discriminant Analysis

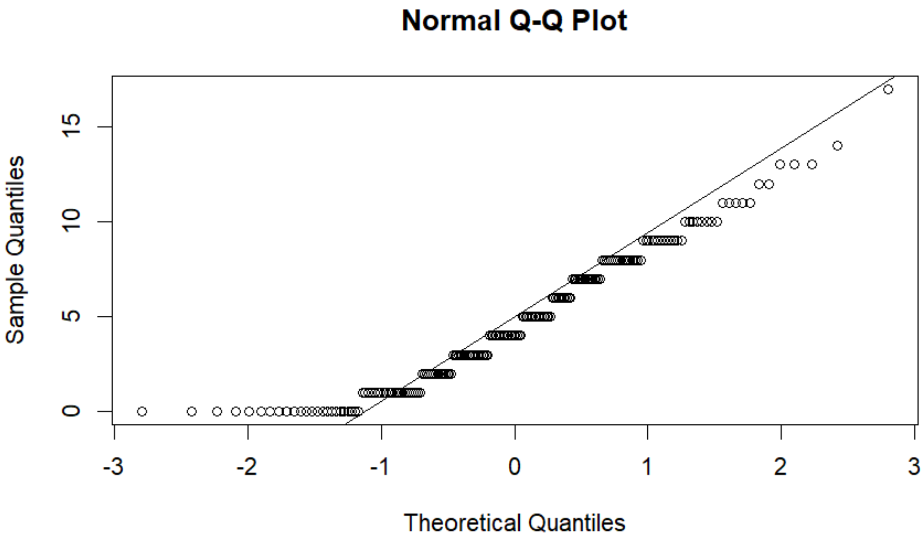
In order to apply LDA and QDA it’s needed to check the following assumptions:

1. Normality of predictors’ conditioned distributions
2. Presence (and removal) of outliers
3. Homoschedasticity
4. Presence (and removal) of multicollinear variables
5. Independency between predictors

Normality check: Pregnancies

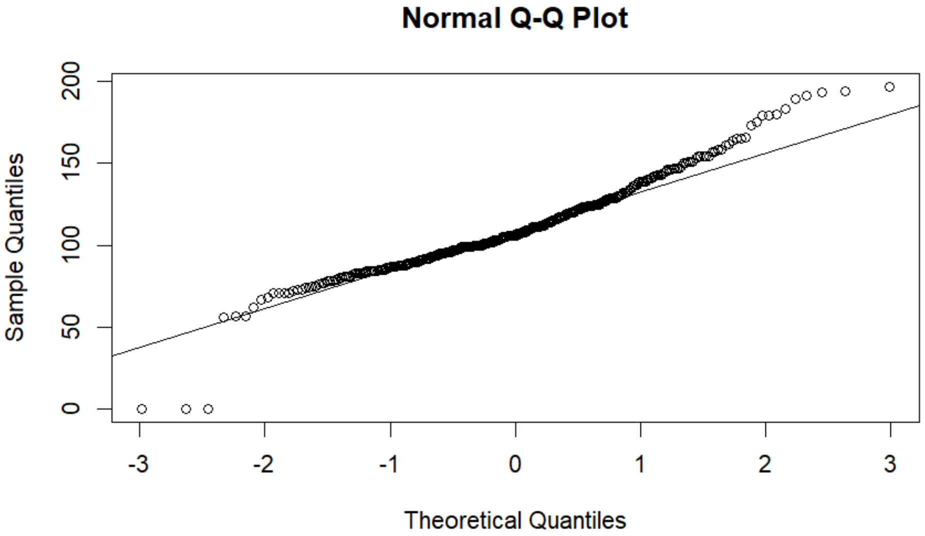


‘Outcome’ class: 0

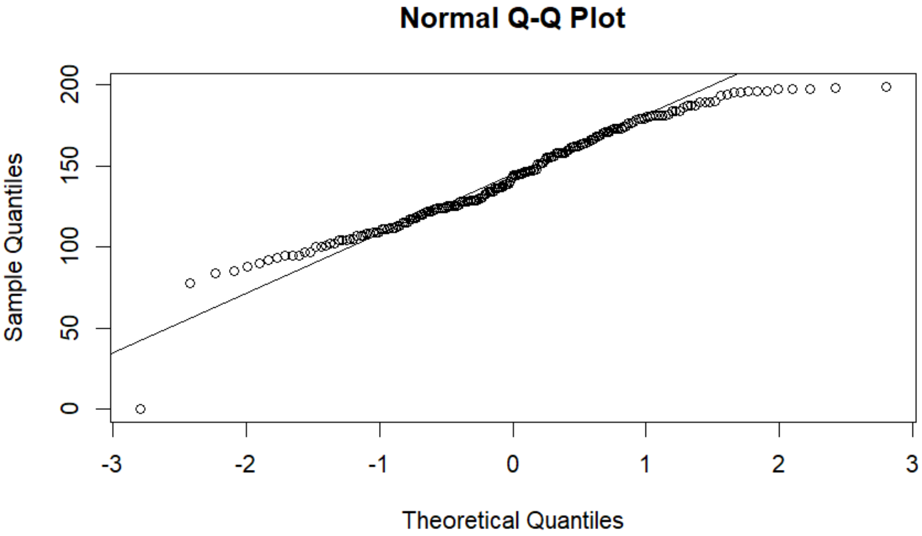


‘Outcome’ class: 1

Normality check: Glucose

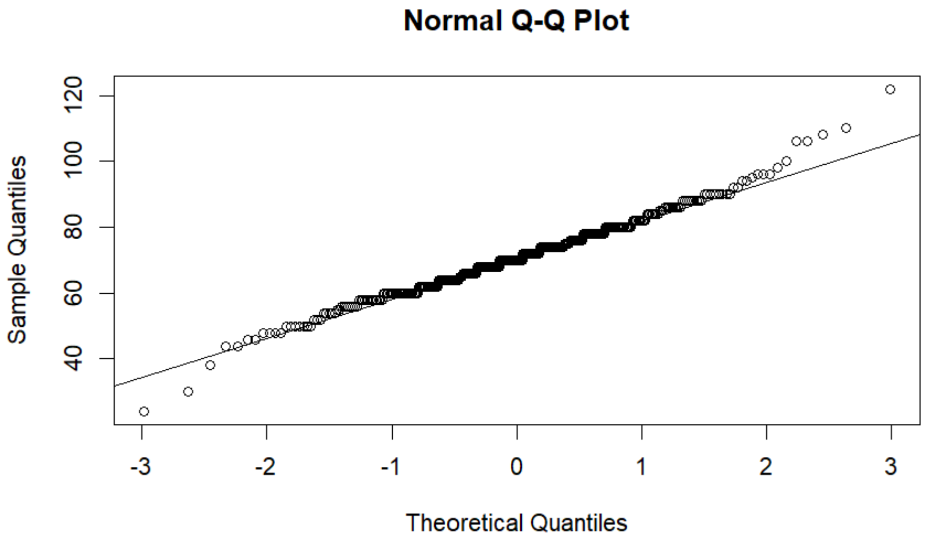


'Outcome' class: 0

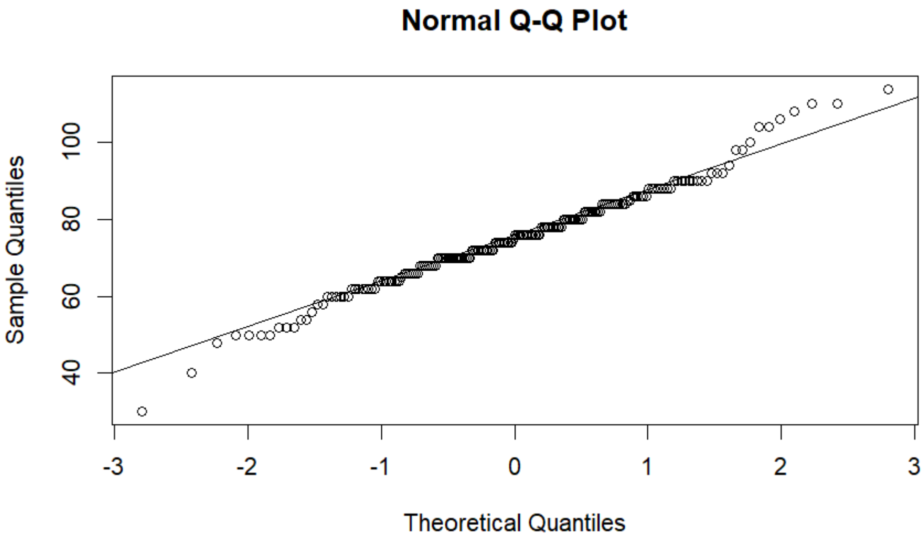


'Outcome' class: 1

Normality check: BloodPressure

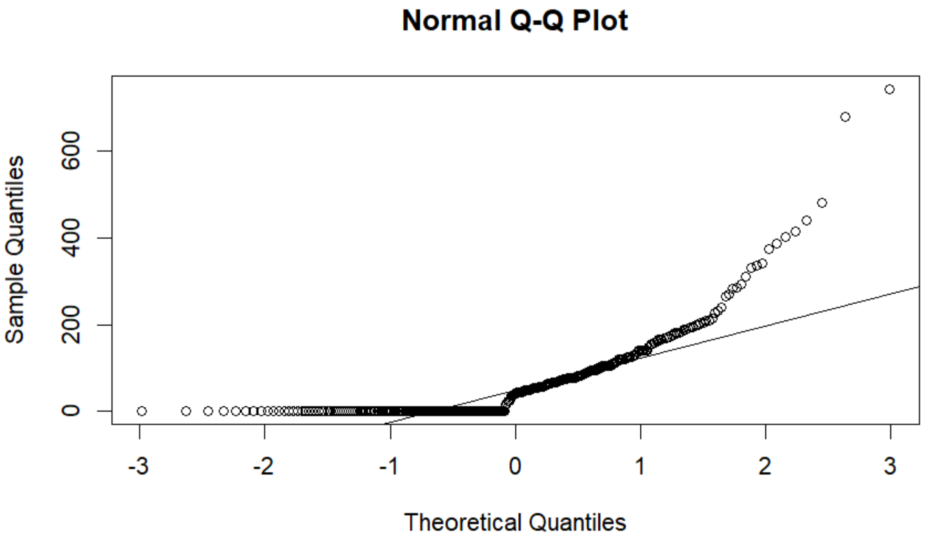


‘Outcome’ class: 0

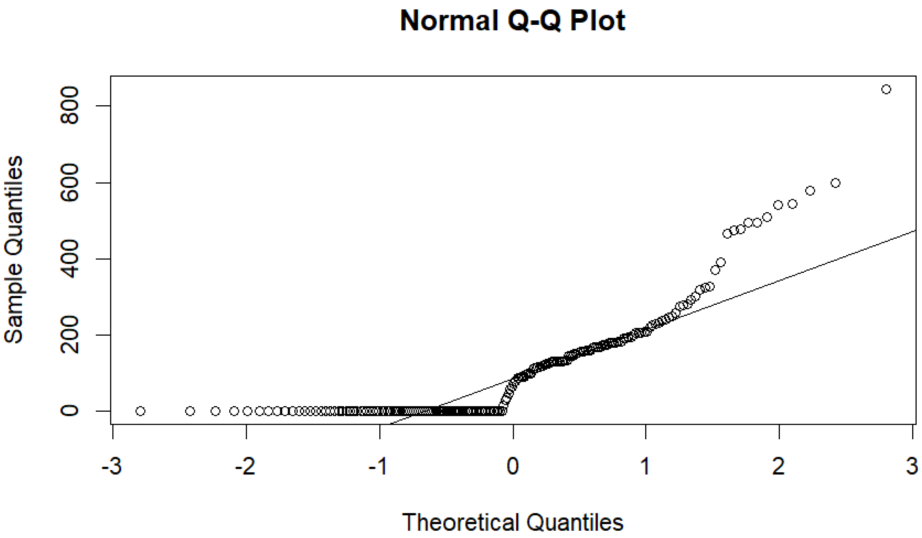


‘Outcome’ class: 1

Normality check: Insulin

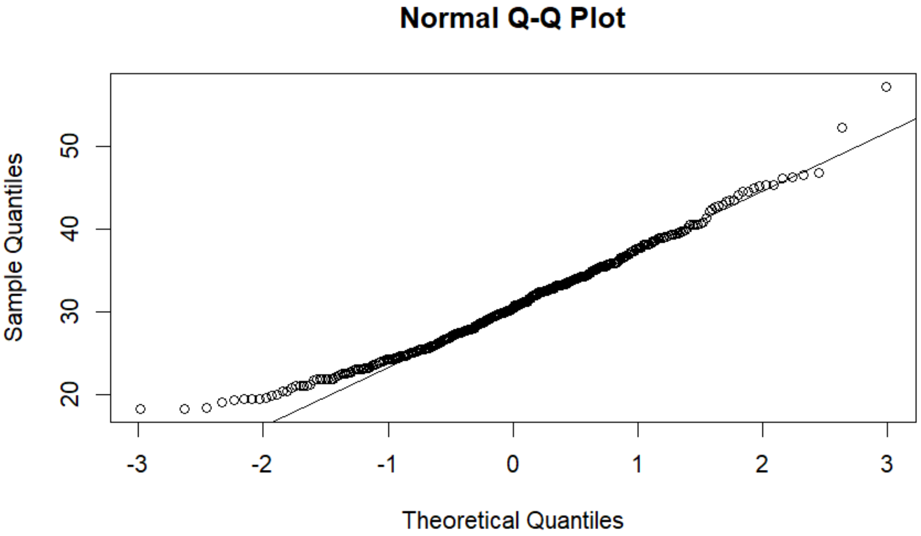


‘Outcome’ class: 0

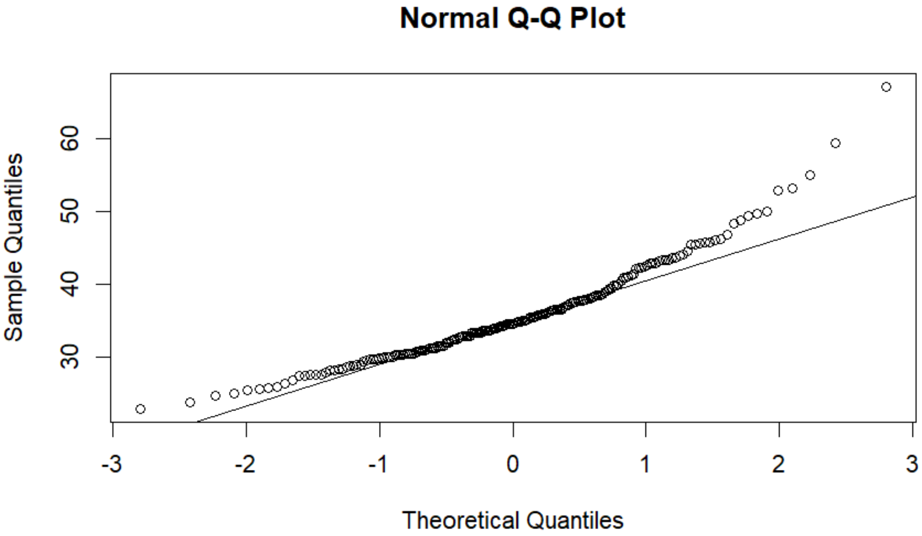


‘Outcome’ class: 1

Normality check: BMI

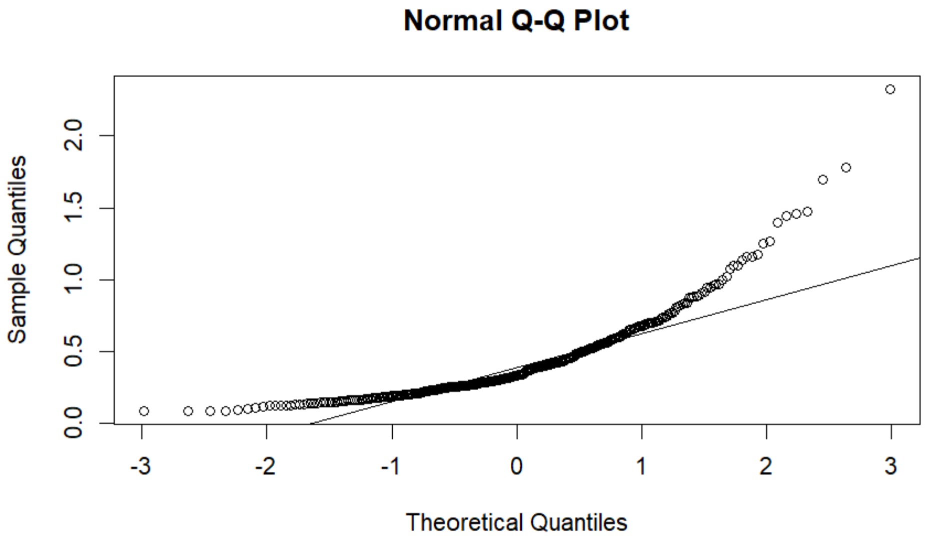


‘Outcome’ class: 0

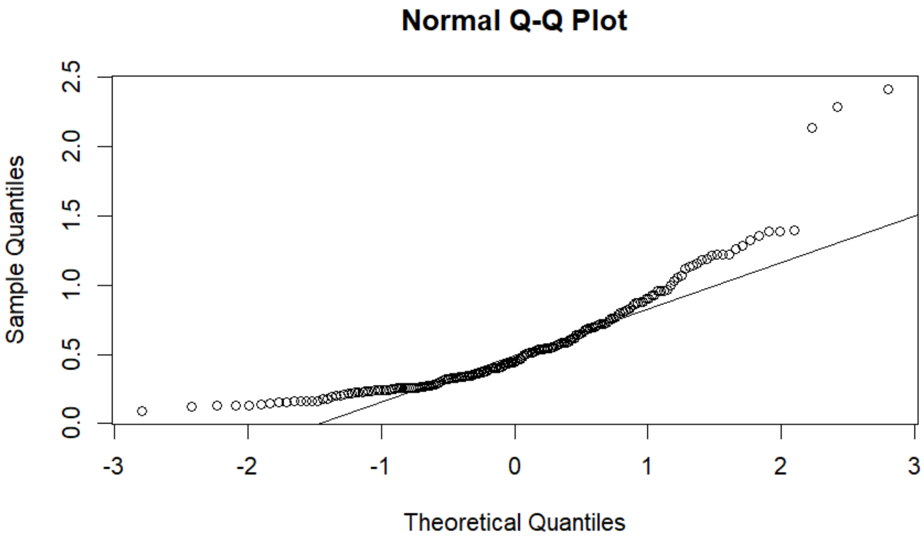


‘Outcome’ class: 1

Normality check: DiabetesPedigreeFunction

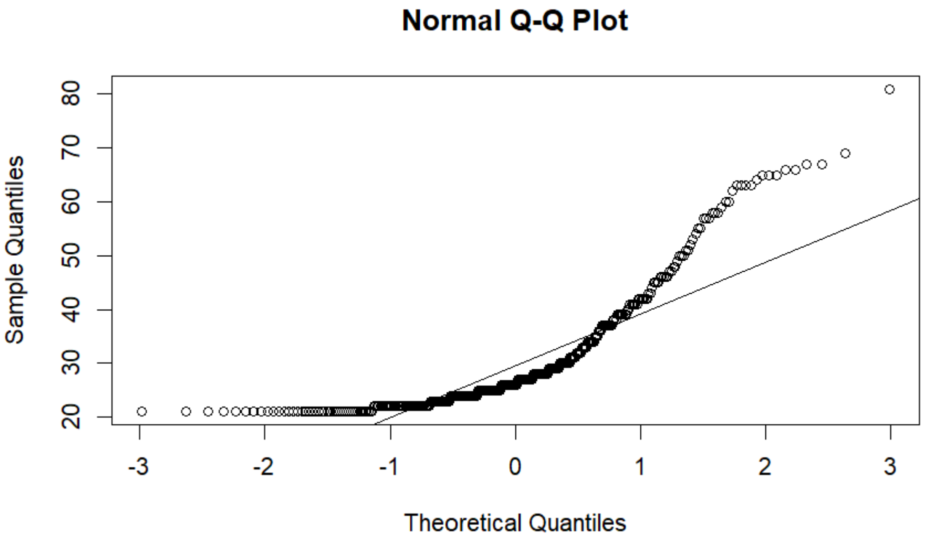


‘Outcome’ class: 0

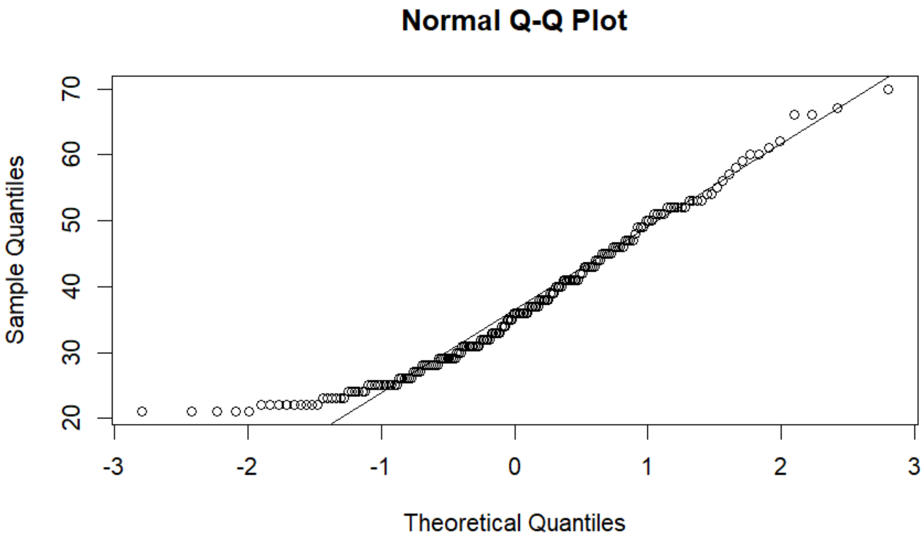


‘Outcome’ class: 1

Normality check: Age



‘Outcome’ class: 0



‘Outcome’ class: 1

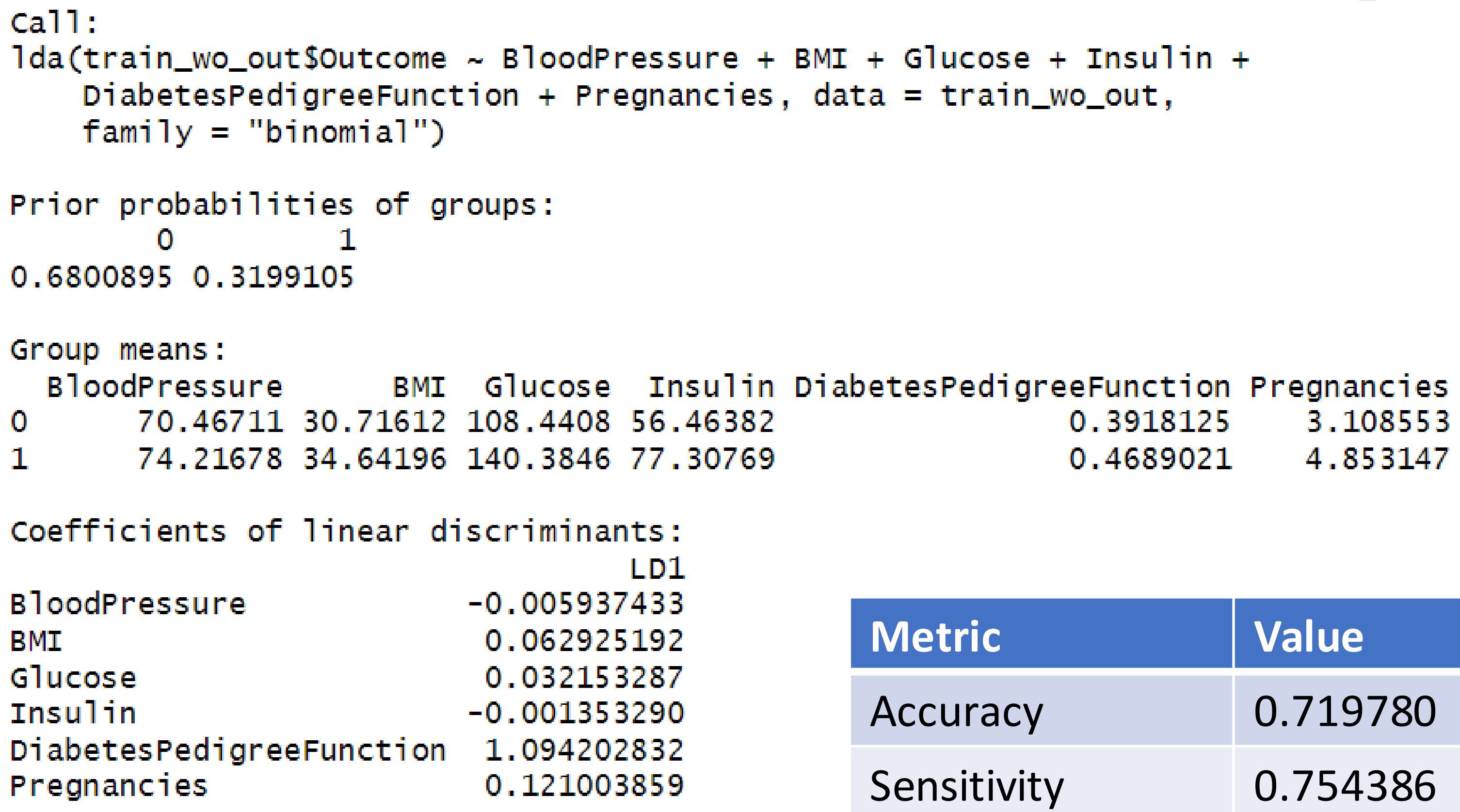
Outliers Detection ad Elimination

Discriminant Analysis is affected by outliers, so it’s needed a check to eliminate them:

Predictors	# outliers
Pregnancies	2
Glucose	4
BloodPressure	14
Insulin	26
BMI	6
DiabetesPedigreeFunction	29
Age	19

Dataset without outliers -> 467

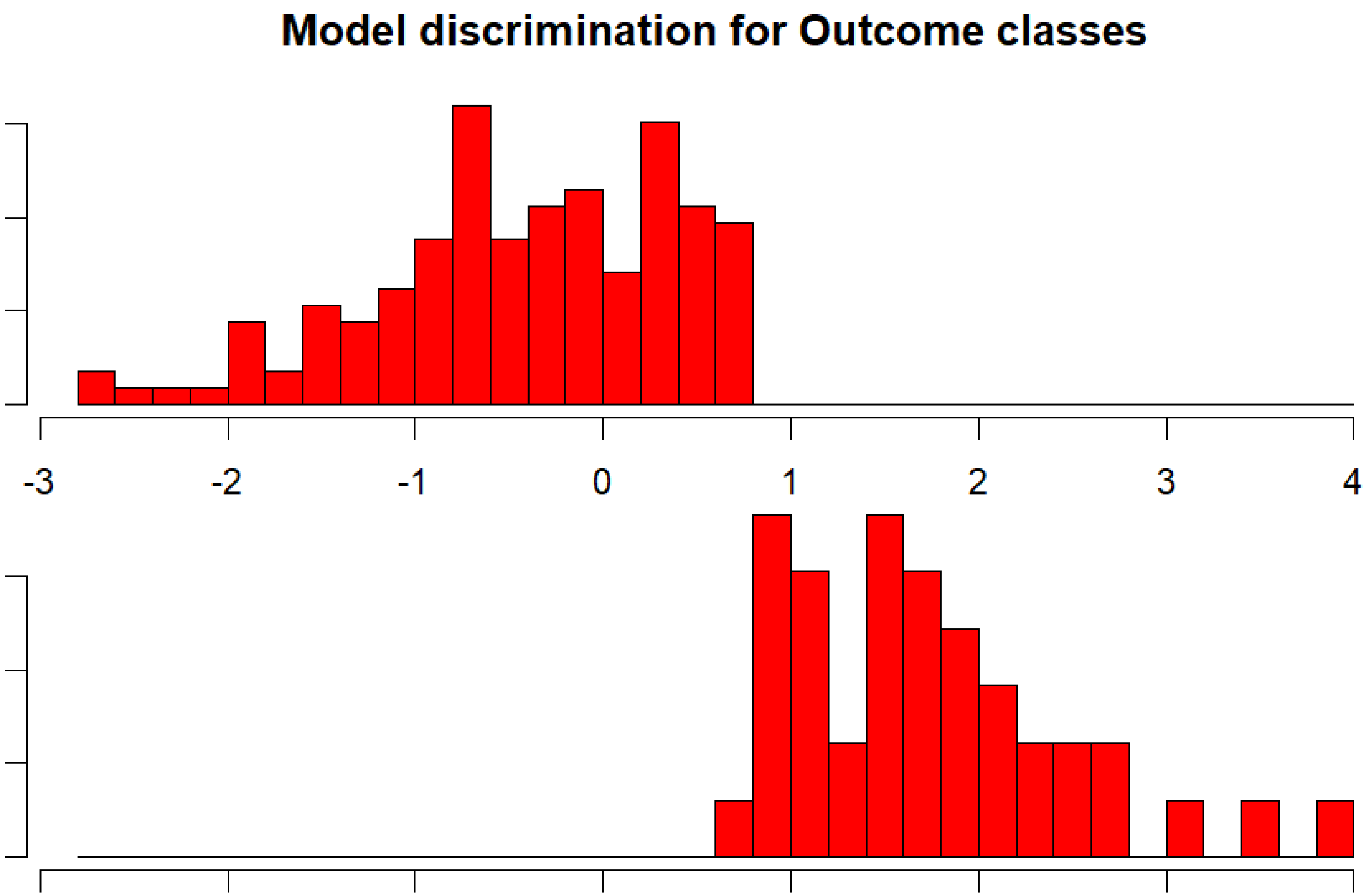
Linear Discriminant Analysis



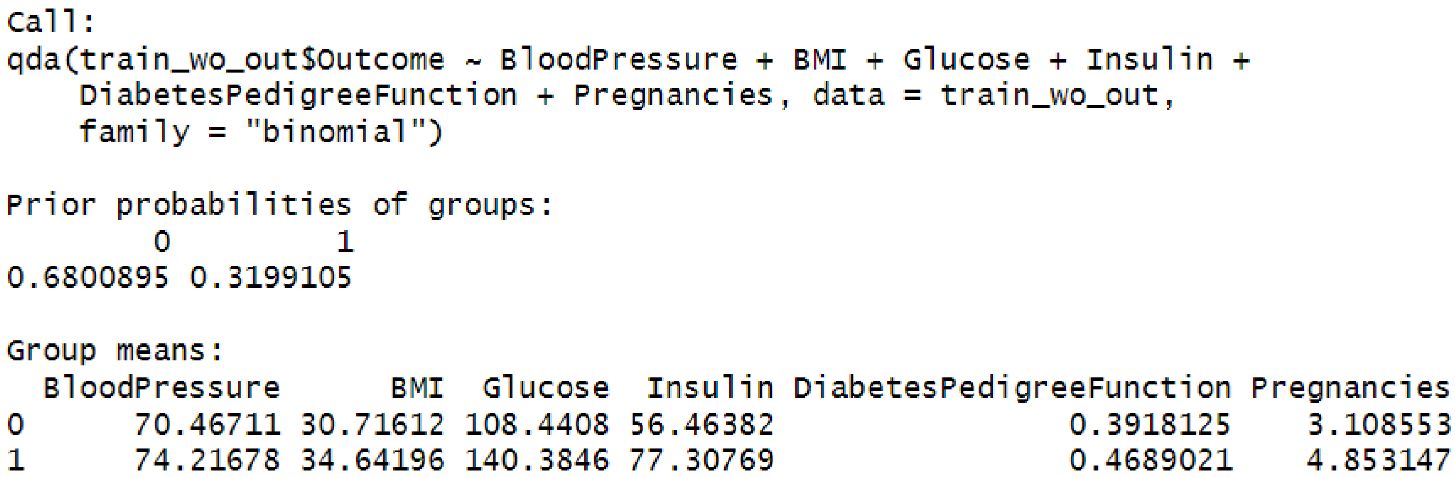
```
Call:
lda(train_wo_out$Outcome ~ BloodPressure + BMI + Glucose + Insulin + 
    DiabetesPedigreeFunction + Pregnancies, data = train_wo_out, 
    family = "binomial")

Prior probabilities of groups:
        0         1 
0.6800895 0.3199105 

Group means:
  BloodPressure      BMI  Glucose  Insulin DiabetesPedigreeFunction Pregnancies
0      70.46711 30.71612 108.4408 56.46382                0.3918125    3.108553
1      74.21678 34.64196 140.3846 77.30769                0.4689021    4.853147

Coefficients of linear discriminants:
                                  LD1
BloodPressure            -0.005937433
BMI                       0.062925192
Glucose                   0.032153287
Insulin                  -0.001353290
DiabetesPedigreeFunction  1.094202832
Pregnancies               0.121003859
```

Metric	Value
Accuracy	0.719780
Sensitivity	0.754386

Model predictions by Outcome classes



Quadratic Discriminant Analysis

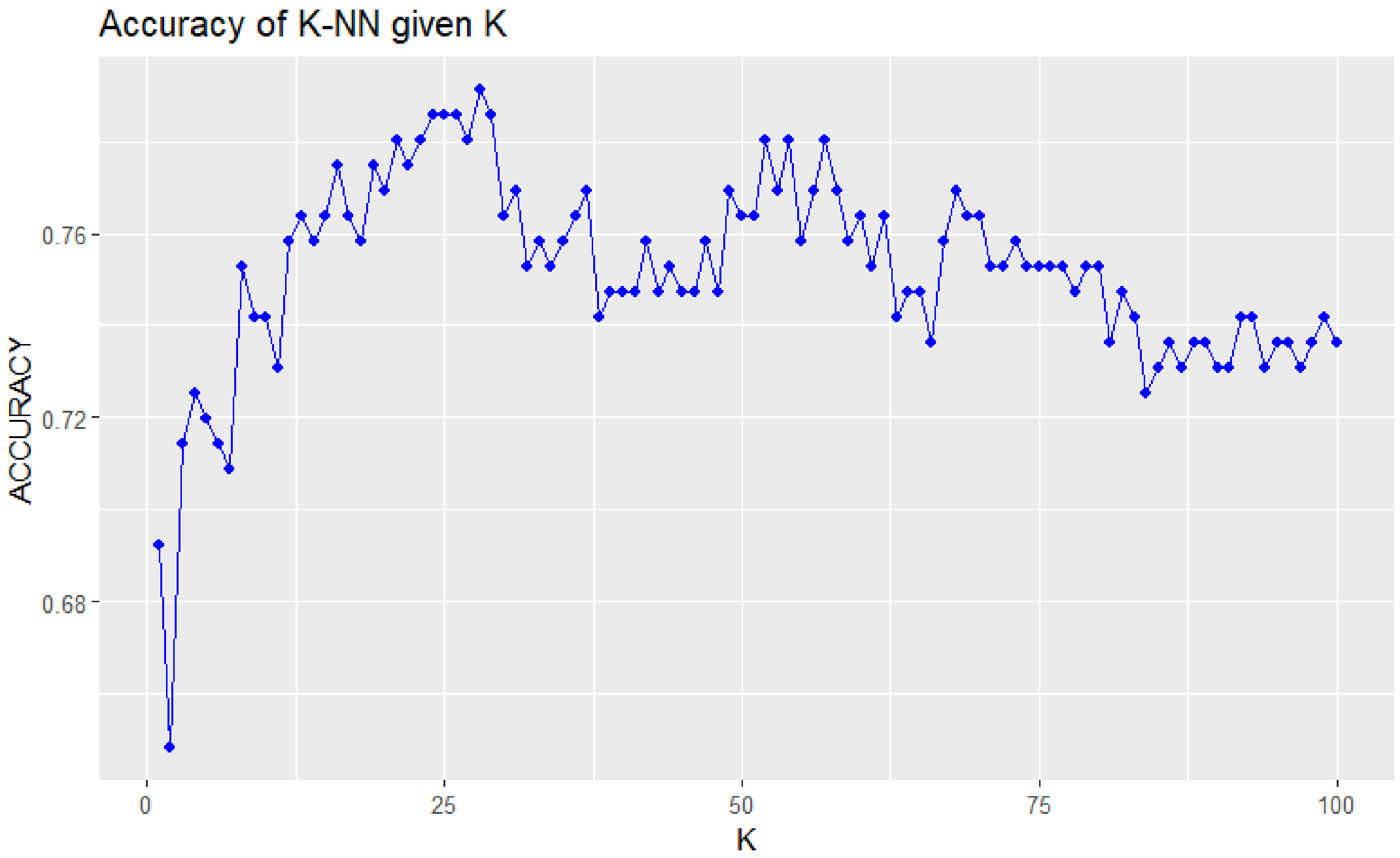


```
Call:
qda(train_wo_out$Outcome ~ BloodPressure + BMI + Glucose + Insulin + 
    DiabetesPedigreeFunction + Pregnancies, data = train_wo_out, 
    family = "binomial")

Prior probabilities of groups:
        0         1 
0.6800895 0.3199105 

Group means:
  BloodPressure      BMI  Glucose  Insulin DiabetesPedigreeFunction Pregnancies
0      70.46711 30.71612 108.4408 56.46382                0.3918125    3.108553
1      74.21678 34.64196 140.3846 77.30769                0.4689021    4.853147
```

Metric	Value
Accuracy	0.6593407
Sensitivity	0.7017544

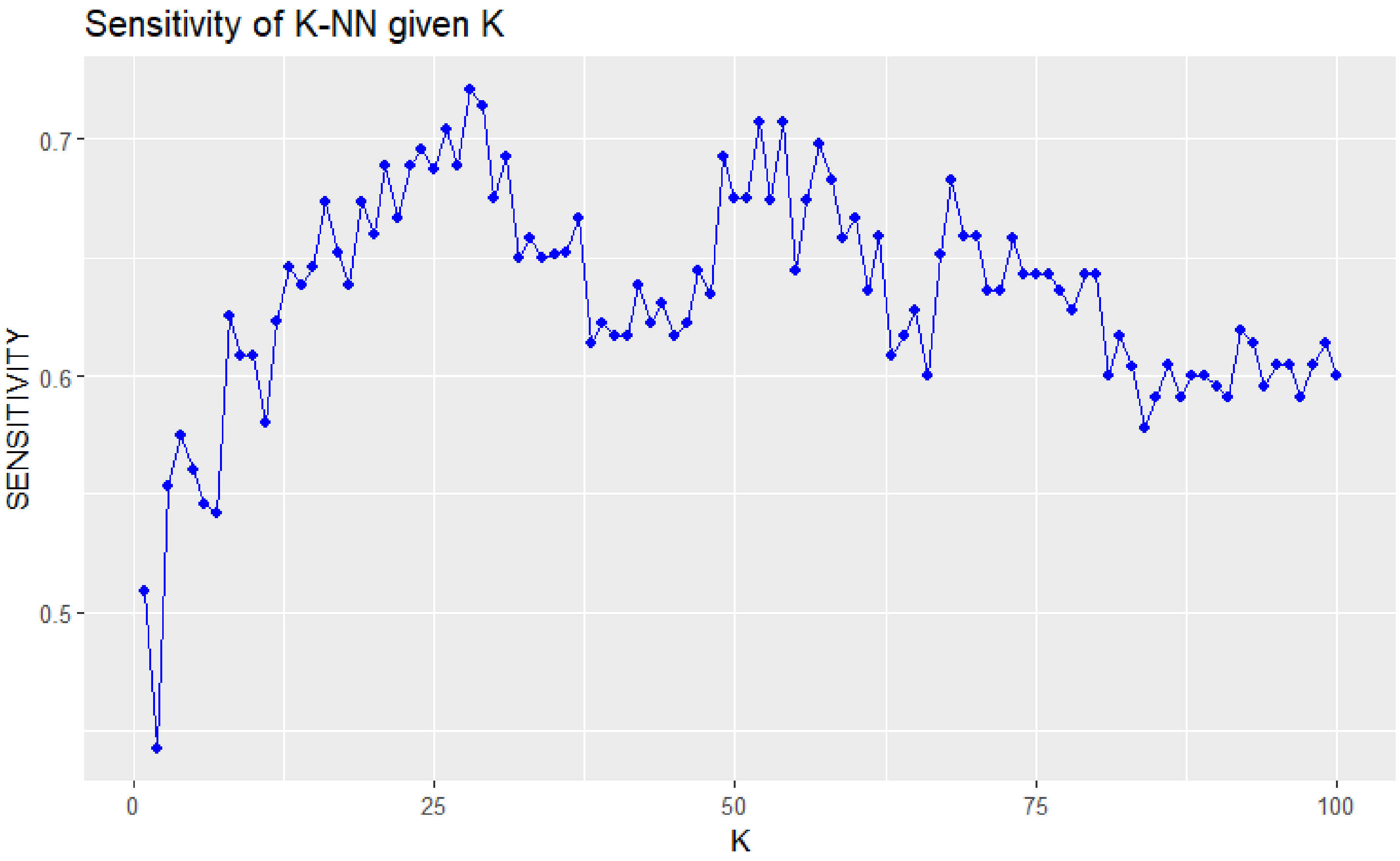
K-NN: k for best accuracy

K=28 produces best accuracy: 0.7912

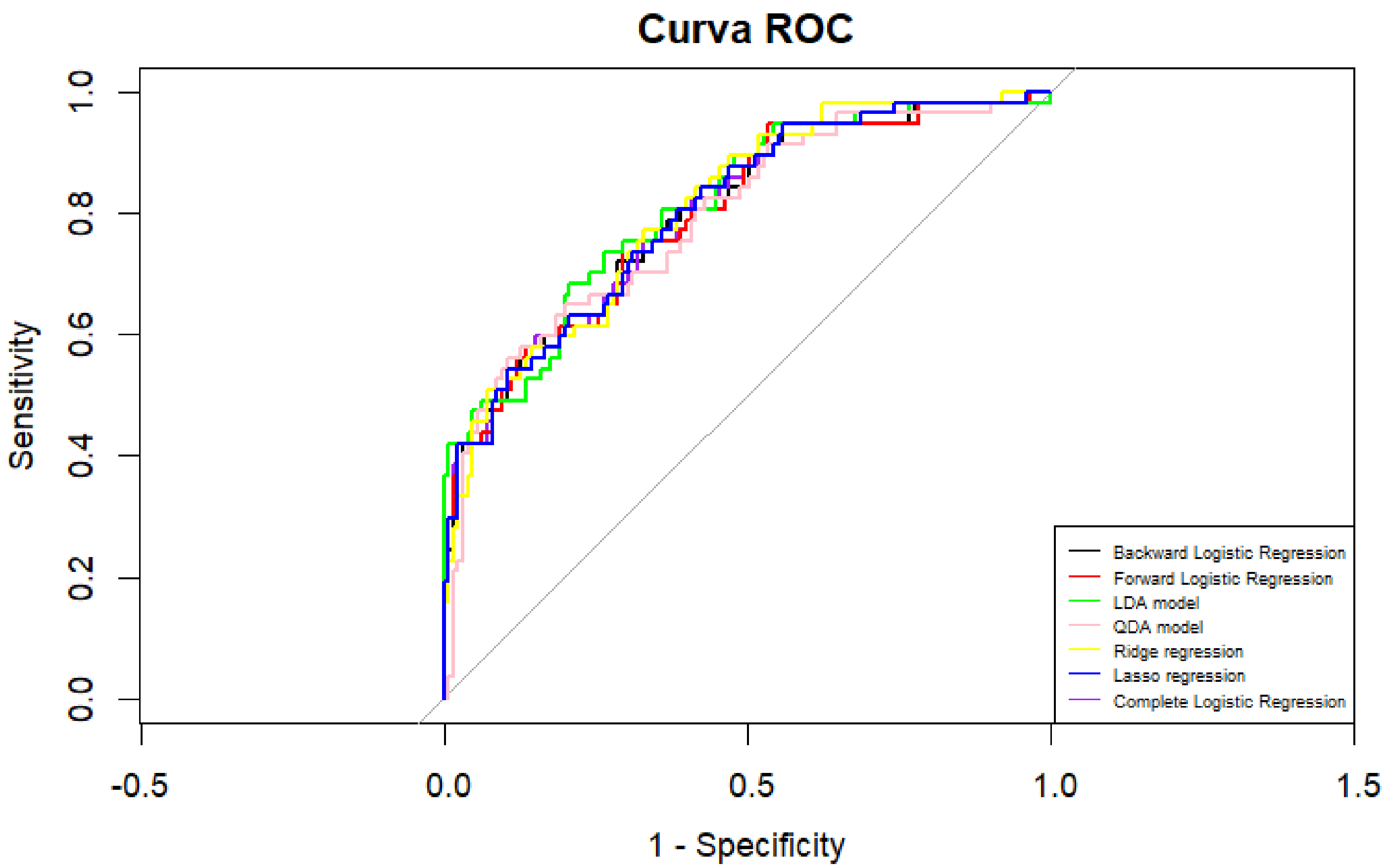


K-NN: k for best sensitivity

K=28 produces even best sensitivity: 0.7209



Model comparison: ROC curve



Model Comparison: Metrics

Model	Accuracy	Sensitivity
Complete Logistic Model	0.6923	0.7368
Backward Logistic Model	0.6923	0.7368
Forward Logistic Model	0.7143	0.7544
Ridge Regression Model	0.6703	0.8246
Lasso Regression Model	0.6868	0.7368
LDA Model	0.7198	0.7544
QDA Model	0.6593	0.7018
28- KNN	0.7912	0.7209

Conclusions

In order to forecast diabetes in out target population, the main gactor to be taken into account are:

- Genetics susceptibility
- Number of pregnancies
- BMI
- Glucose level